GRiT: A Generative Region-to-text Transformer for Object Understanding

Year: 2022 | Citations: 144 | Authors: Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu

---

Abstract

This paper presents a Generative RegIon-to-Text transformer, GRiT, for object understanding. The spirit of GRiT is to formulate object understanding aspairs, where region locates objects and text describes objects. For example, the text in object detection denotes class names while that in dense captioning refers to descriptive sentences. Specifically, GRiT consists of a visual encoder to extract image features, a foreground object extractor to localize objects, and a text decoder to generate open-set object descriptions. With the same model architecture, GRiT can understand objects via not only simple nouns, but also rich descriptive sentences including object attributes or actions. Experimentally, we apply GRiT to object detection and dense captioning tasks. GRiT achieves 60.4 AP on COCO 2017 test-dev for object detection and 15.5 mAP on Visual Genome for dense captioning. Code is available at https://github.com/JialianW/GRiT